# A Time-to-Event Analysis of Heart Failure via Electronic Health Records

Deesha Bhaumik[1], Christian Erickson[2] & Tahmeed Tureen[3]
[1]Johns Hopkins University, [2] Hope College, [3] University of Michigan – Ann Arbor

## Introduction

Chronic heart failure is a major and increasingly common medical issue today, accounting for over one million hospitalizations each year. As the world of statistics migrates to the big data era, large and complex data sets such as Electronic Health Records (EHR) are becoming readily available for scientific research. With the combination of EHR and statistics, the potential of drawing inference on hidden relationships between several diseases and health outcomes is growing. In this research project, we conducted a time-to-event analysis on the EHR data set from the Michigan Genomics Initiative (MGI) to create a heart failure prediction model using scientifically acknowledged risk factors.

## Overview

**Scientific Motivation**

To create a statistical model from EHR that can potentially inform patients if and by how much they are at risk for heart failure at each diagnosis. This could go a long way in preventing heart failure in the future.

**Data Wrangling**

- Integration of data with ~3,000,000 observations
- R Packages: ggplot2, dplyr, tidyr
- Final dataset with 9133 observations for survival analysis

**Challenges**

- Electronic Health Records
  - Not collected for research purposes
  - Textual data differences in diagnosis
  - ICD-9 and ICD-10 Diagnosis Codes
  - Traditional Missing Data

**Missing Data**

- All diagnostic codes that did not have a corresponding time of diagnosis were dropped prior to construction of the analytic dataset

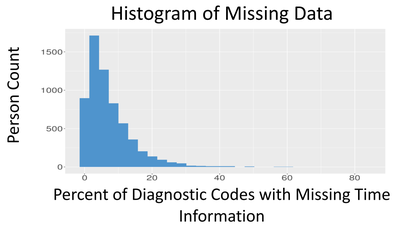


**Figure 1.** Plots the distribution of un-timed diagnostic codes by patient

## Cohort Definition

- A study cohort free of heart failure from the MGI EHR was carefully constructed by defining a baseline at age 40 or older
- A year's worth of data prior to baseline was used to determine status of specified risk factors
- Our outcome was Heart Failure (ICD 9: 428, ICD 10: I50)

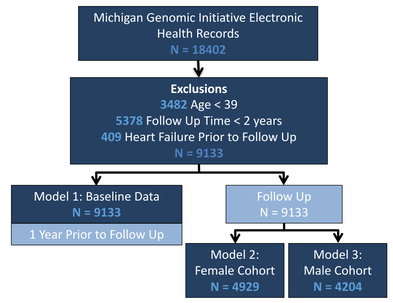


**Figure 2.** Study Schema

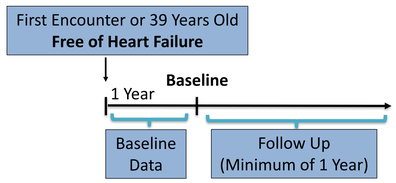


**Figure 3.** Cohort Selection

| Risk Factor | Female Percentage (mean) | Male Percentage (mean) |
|---|---|---|
| **Age** | (53.06) | (57.16) |
| **Arrhythmia** | 1.907 | 2.807 |
| **Cardiomyopathy** | 0.243 | 0.785 |
| **Emphysema** | 0.791 | 1.261 |
| **Failure Chronic Kidney Disease** | 0.142 | 0.547 |
| **Hepatitis C** | 0.264 | 0.452 |
| **Hypercholesterol** | 1.887 | 3.687 |
| **Hypertension** | 15.054 | 21.908 |
| **Mild Chronic Kidney Disease** | 0.162 | 0.143 |
| **Moderate Chronic Kidney Disease (CKD)** | 0.467 | 0.714 |
| **Morbid Obesity** | 3.023 | 1.689 |
| **Severe Chronic Kidney Disease** | 0.284 | 0.642 |
| **Sleep Disorder** | 3.713 | 3.949 |
| **Type 2 Diabetes Mellitus** | 5.823 | 7.826 |
| **Type 2 Diabetes Mellitus*Age** | (3.192) | (4.652) |

**Table 1.** Study Cohort Summary Statistics

## Survival Analysis

- Cox Proportional Hazards Regression (Coxph) was run to conduct a survival analysis on our defined EHR cohorts
- Covariates (Sex, Type II Diabetes) were manipulated to address the proportionality assumptions

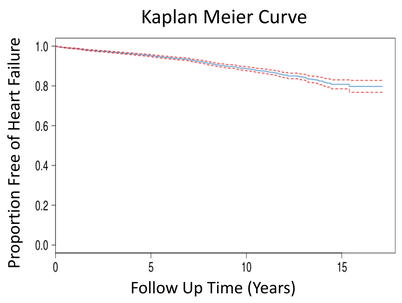


**Figure 4.** Empirical Kaplan Meier Curve

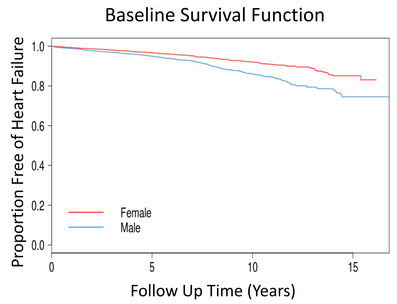


**Figure 5.** Baseline Survival Function Curves comparing Female and Male

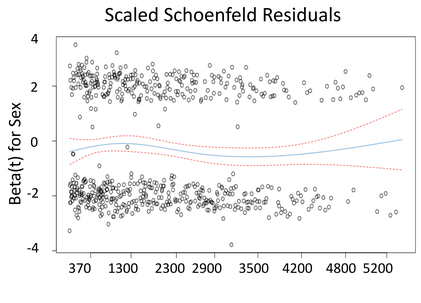


**Figure 6.** Original Violation of the Cox Model Assumption for Sex Covariate

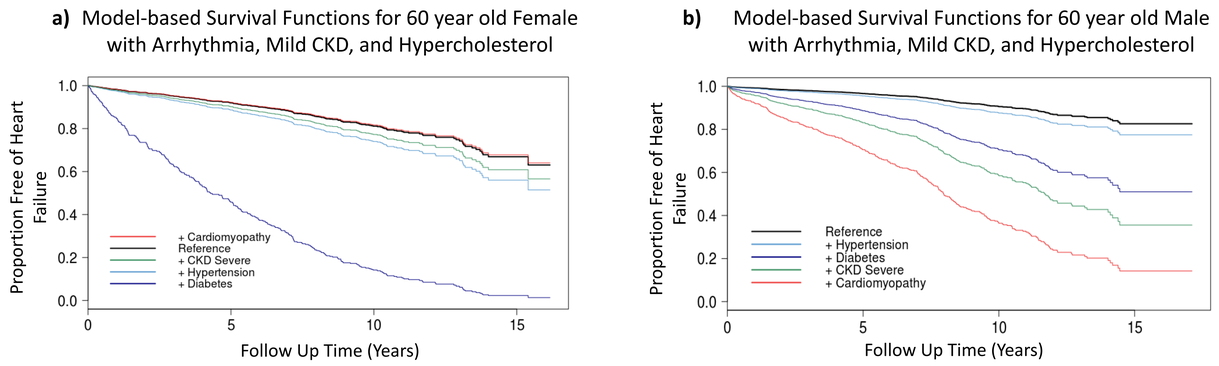


**Figure 7 (a-b).** Kaplan Meier Curves: Survival functions for reference person with characteristics of age 60 with arrhythmia, mild CKD, and hypercholesterol comparing (a) Female and (b) Male when adding additional risk factors

## Results

**Table 2.** Fitted Effect Sizes for Female Cohort

| Female (n = 4929) Heart Failure (n = 270) | | | | |
|---|---|---|---|---|
| **Risk Factor** | **n (mean)** | **HR** | **P-Value** | **95% CI** |
| Age | (53.06) | 1.052 | **4.44E-16** | (1.040, 1.065) |
| Type 2 Diabetes Mellitus | 287 | 9.350 | **0.020** | (1.433, 61.024) |
| Hypertension | 742 | 1.442 | **0.027** | (1.043, 1.993) |
| Sleep Disorder | 183 | 1.764 | **0.044** | (1.015, 3.068) |
| Failure Chronic Kidney Disease | 7 | 4.179 | 0.058 | (0.954, 18.305) |
| Type 2 Diabetes Mellitus*Age | (3.192) | 0.973 | 0.105 | (0.942, 1.006) |
| Arrhythmia | 94 | 1.563 | 0.139 | (0.865, 2.825) |
| Morbid Obesity | 149 | 1.635 | 0.158 | (0.826, 3.234) |
| Hepatitis C | 13 | 3.474 | 0.215 | (0.485, 24.904) |
| Moderate Chronic Kidney Disease | 23 | 2.169 | 0.225 | (0.621, 7.578) |
| Severe Chronic Kidney Disease | 14 | 2.070 | 0.342 | (0.462, 9.268) |
| Hypercholesterol | 93 | 0.776 | 0.563 | (0.328, 1.836) |
| Mild Chronic Kidney Disease | 8 | 1.675 | 0.607 | (0.234, 11.984) |
| Emphysema | 39 | 1.333 | 0.689 | (0.326, 5.445) |
| Cardiomyopathy | 12 | 0.967 | 0.974 | (0.131, 7.166) |

**Table 3.** Fitted Effect Sizes for Male Cohort

| Male (n = 4204) Heart Failure (n = 363) | | | | |
|---|---|---|---|---|
| **Risk Factor** | **n (mean)** | **HR** | **P-value** | **95% CI** |
| Age | (57.16) | 1.052 | **<2e-16** | (1.041, 1.063) |
| Cardiomyopathy | 33 | 10.203 | **<2e-16** | (6.269, 16.606) |
| Hypertension | 921 | 1.337 | **0.036** | (1.020, 1.753) |
| Severe Chronic Kidney Disease | 27 | 3.080 | **0.039** | (1.059, 8.961) |
| Type 2 Diabetes Mellitus | 329 | 3.526 | 0.227 | (0.456, 27.272) |
| Emphysema | 53 | 1.684 | 0.251 | (0.6914, 4.101) |
| Moderate Chronic Kidney Disease | 30 | 1.725 | 0.290 | (0.628, 4.734) |
| Hepatitis C | 19 | 1.945 | 0.352 | (0.490, 7.887) |
| Arrhythmia | 118 | 1.259 | 0.372 | (0.759, 2.089) |
| Type 2 Diabetes Mellitus*Age | (4.652) | 0.987 | 0.447 | (0.956, 1.020) |
| Mild Chronic Kidney Disease | 6 | 0.569 | 0.601 | (0.069, 4.704) |
| Hypercholesterol | 155 | 0.911 | 0.738 | (0.526, 1.576) |
| Failure Chronic Kidney Disease | 23 | 1.191 | 0.802 | (0.304, 4.664) |
| Sleep Disorder | 166 | 1.036 | 0.915 | (0.545, 1.966) |
| Morbid Obesity | 71 | 1.064 | 0.916 | (0.334, 3.392) |

## Discussion

**Present**

- Statistically Significant Risk Factors include:
  - Women: Age, Type 2 Diabetes Mellitus, Hypertension, and Sleep Disorder
  - Male: Age, Cardiomyopathy, Hypertension, and Severe CKD
- Significant risk factors are consistent with some of the claims made by scientific journals from NIH about certain heart failure risk factors

**Future**

- Further analysis to assess prediction accuracy
- Exploring shrinkage methods to carefully analyze the hazards ratio of "big impact" risk factors
- Future steps may include analysis of risk factors and understanding how to prevent their development
- Better understanding and discussing the potential of Electronic Health Records in improving scientific research

## Contact Information

Deesha Bhaumik [ Email: dbhaumi1@jhu.edu ]
Christian Erickson [ Email: christian.erickson@hope.edu ]
Tahmeed Tureen [ Email: tureen@umich.edu, Website: Link, GitHub: Link ]

## Acknowledgements

Bhramar Mukherjee, PhD
Phil Boonstra, PhD
Zhenke Wu, PhD
Matthew Zawistowski, PhD
Xutong Zhao, PhD Candidate
Department of Biostatistics, University of Michigan